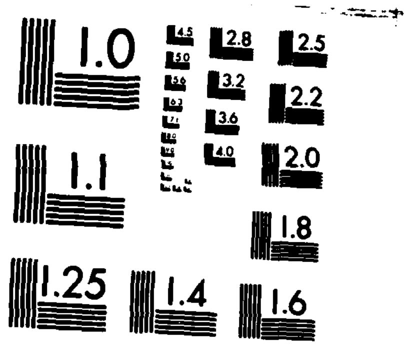
AD-A153 282 A STUDY OF ERROR RATES IN VOICE RECOGNITION(U) AIR FORCE AEROSPACE MEDICAL RESEARCH LAB WRIGHT-PATTERSON AFB OH J S BALLMANN FEB 85 AFAMRL-TP-85-300 1/1

UNCLASSIFIED F/G 17/2 NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD-A153 282

AFAMRL-TR-85-300

A STUDY OF ERROR RATES IN VOICE RECOGNITION

JENNIFER S. BALLMAN

AIR FORCE AEROSPACE MEDICAL RESEARCH LABORATORY

FEBRUARY 1985

DTIC FILE COPY

Approved for public release; distribution unlimited.

AIR FORCE AEROSPACE MEDICAL RESEARCH LABORATORY
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from Air Force Aerospace Medical Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
Cameron Station
Alexandria, Virginia 22314

TECHNICAL REVIEW AND APPROVAL

AFAMRL-TP-85-300

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

The voluntary informed consent of the subjects used in this research was obtained as required by Air Force Regulation 169-3.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

CHARLES BATES, JR.
Director, Human Engineering Division
Air Force Aerospace Medical Research Laboratory

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): AFAMRL-TP-85-300	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: AFSC, AMD, AF Aerospace Medical Research Laboratory — 6b. OFFICE SYMBOL (If applicable): HEC	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State and ZIP Code): Wright-Patterson AFB OH 45433-6573	7b. ADDRESS (City, State and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION — 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State and ZIP Code)	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO
62202F	7184	27	03

11. TITLE (Include Security Classification) A STUDY OF ERROR RATES IN VOICE RECOGNITION (U)

12. PERSONAL AUTHOR(S) Ballmann, Jennifer S.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Technical	FROM ____ TO ____	February 1985	15

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Voice recognition, command and control, voice control, computer-human interaction,

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Voice recognition as a means of data entry was evaluated by measuring error rates of words in a test vocabulary. When the words were divided into groups based on certain phonetic characteristics, significant differences in error rate were found between the groups. Two rules were proposed as an aid in selecting commands to system designers contemplating use of voice recognition.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☒ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Capt David G. Leupp	(513) 255-7591	AFAMRL/HEC

PREFACE

This research was conducted while the author was a University of Dayton student cooperative at the Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio. The author gratefully acknowledges the support and contributions of Ms. Sharon Ward, Capt David Leupp, and MSgt Danny Bridges.

INTRODUCTION

Complex man-machine systems are often limited by comparatively slow and inefficient data entry methods. The recent proliferation of computer peripherals, such as trackballs, "mice", touch screen displays, digitizing panels, and voice command systems attests to this problem. The most complex solution, voice recognition, is also the most attractive because speech is the richest and most natural way for people to communicate. Machines are available that decipher phonetic patterns into text with over 90% accuracy, but in critical situations, even a small number of errors can significantly degrade performance. It is therefore of interest to identify words that are consistently unintelligible or mistaken for other words, and avoid their use in voice command systems. The Data Entry Vocabularies in a C^3 Environment study (DEVICE) was performed during December 1983-January 1984 and used a vocabulary from a prior AFAMRL experiment to investigate error rates. The results are presented here.

BACKGROUND

During the SIMCOPE-1 study, performed by Dr. Peter Crane (Univ. of Pittsburgh) and Capt Dave Leupp (AFAMRL/HEC)[1], which incorporated voice command as one of two data entry methods in a simulated missile warning crewstation, several "problem" words emerged. A pilot study to find word recognition error rates for the entire SIMCOPE-1 vocabulary produced more candidate words. Replacement words with meanings similar to the problem words were chosen for phonetic dissimilarity and combined with the original list of 95 words, giving a total of 125 words.

EXPERIMENTAL DESIGN

The experimental vocabulary contains three groups of words (Table 1), the control group (words which caused few errors, totalling 65), the original group of problem words (totalling 30), and the replacement group (totalling 30). Two vocabularies were used: control and original (V1), and control and replacement (V2). Ten subjects from a subject pool were trained on the voice recognition equipment. All 125 words were repeated ten times to allow the machine to "learn" the pronunciation of the word, and then this information was stored on tape. The system used was speaker-dependent, requiring a different tape for each subject. Training was supervised by the experimenter, who coached the subjects to avoid monotonous pronunciation. (Since word inflection is invariably different during training than during use, the processor, which averages the training pronunciations, will recognize words better if various inflections are used during training.)

During a session, a subject was seated in a soundproof room in front of a terminal, wearing a head-mounted microphone. Microphone placement has been shown to be an important factor in recognition,[2,3] so placement was supervised during training and trials. The trial was initiated by the subject and consisted of three randomly ordered iterations of V1 or V2, with each word flashing onto the screen at random intervals (1-3 sec) and remaining on the screen for .35 sec. The subject attempted to read the word into the microphone before the next word appeared. Time stress was present to simulate a more realistic setting, and to prevent the subject

from lapsing into a monotone (yet few subjects skipped words because of it). Subjects could interrupt the experiment at any time and had two mandatory breaks per trial. Sessions lasted approximately 25 minutes, consisting of two trials separated by a five-minute break, for a total of five breaks. Each session included one trial of V1 and one of V2 in varying order. The session order for subjects is shown in Table 2. Audio tapes of the sessions were made.

Equipment used included a Threshold 600 voice recognizing unit, a Shure SM10A head-mounted microphone worn by the subjects, a DEC PDP 11/40 minicomputer that presented trials and collected data, and a Tascam 44 audio tape recorder.

RESULTS

Two types of errors were recorded, misrecognition, or confusion with another word, and nonrecognition, or failure of the system to match the word with the training pattern. Nonrecognitions can be frustrating to a user (especially with the usual audible feedback), but misrecognitions are more dangerous to system performance because they can go undetected.

The error rates for each group are shown in Table 3. It is clear that the intuitive criteria used to select the replacement group did not result in a superior vocabulary. In fact, the replacement words had a significantly higher misrecognition rate ($x^2 = 6.00$, $p < .05$). One possible reason might be that many of the subjects, drawn from a limited pool, had unavoidedly been subjects for the SIMCOPE-1 study and were more familiar with the original words (V1). Table 4 shows the word replacement pairs, each having one original and one replacement word, for which error rates differed significantly. The overall error rate of the best words from each pair was 6.4%, compared with 18.5% for the worst words.

A closer examination of Table 4 and the least recognized words (Table 5) allows some hypotheses to be made based on phonetic qualities of "problem" words:

1. Monosyllabic words are less often recognized than polysyllabic words.

2. Words ending with T or containing a T which is slurred or absent in normal speech (eigh<u>t</u>, wes<u>t</u>, del<u>t</u>a) are also poorly recognized.

Table 6 illustrates the breakdown of the vocabulary into these two groups. The actual number of high-error words in a phonetic group was compared with the expected number, or the number of words in a group multiplied by the overall probability of error over 10% (39/125) (Table 7). Vocabulary words that belonged to either of these groups were significantly more likely to have an overall error rate greater than 10%. Words that belonged to both groups made too small a sample to have a significant difference in rate. It is clear that these word groups should be avoided by system designers especially because words belonging to neither group were highly unlikely to have an error rate over 10%.

TABLE 1. DEVICE Word List

CONTROL

YES	INDISTINCT	EVENT MESSAGES
NO	CDC	SYSTEM REPORTS
NORTH	CWC	TELEPHONE DIRECTORY
SOUTH	BSS	DETAIL MAP
CENTRAL	KEY NORTH	EVENT TIMELINE
INN	NORTH CITY	INTELLIGENCE REPORTS
OUTT	TOLL CITY	OUTPUT FORMAT
WEST	HAYES	REFERENCE DIRECTORY
SUSPECTED	CLEAR	ADS1
ZERO	ASSIGN	ADS2
ONE	BACKSTEP	CLEAR ENTRY
TWO	AUTO	ADS
THREE	EDIT	KNOWN SITES
FOUR	WHITE SANDS	MILITARY INSTALLATIONS
FIVE	PINE GROVE	INDUSTRIAL CENTERS
SIX	SOUTHRICH	ALL EVENTS
SEVEN	HOSTILE	ALL
EIGHT	TEST	ADS GSF
NINE	REASSIGN	BSS GSF
ENTER	SHOW	OCEAN CITY
TYPE I	SUPPRESS	VECTOR
TYPE II	SITUATION MAP	

ORIGINAL

KNOWN	UP ARROW	E7
UNKNOWN	DOWN ARROW	E8
FINISH	LEFT ARROW	E9
BURF	RIGHT ARROW	E10
RIVERTON	E1	NOT CLEAR
LIVINGSTON	E2	LOCATE
DELTA	E3	LOG
SOUTHERN	E4	SUSPECT SITES
SEND	E5	DELETE
ACKNOWLEDGE	E6	FAN

REPLACEMENT

IDENTIFIED	SCROLL UP	EVENT 7
UNIDENTIFIED	SCROLL DOWN	EVENT 8
OVER	SCROLL LEFT	EVENT 9
BRF	SCROLL RIGHT	EVENT 10
ROSEDALE	EVENT 1	UNRESOLVED
LAKEVIEW	EVENT 2	COORDINATES
DAIRYLAND	EVENT 3	MESSAGE LOG
MOUNTAIN	EVENT 4	POSSIBLE SITES
SUBMIT	EVENT 5	REMOVE
OK	EVENT 6	RANGE

TABLE 2. Experimental Design

SUBJECT #	TRIAL	SESSION 1	2	3	4	5	6
1	1	V1	V2	V1	V2	V1	V2
	2	V2	V1	V2	V1	V2	V1
2	1	V2	V1	V2	V1	V2	V1
	2	V1	V2	V1	V2	V1	V2
3	1	V1	V2	V1	V2	V1	V2
	2	V2	V1	V2	V1	V2	V1
4	1	V2	V1	V2	V1	V2	V1
	2	V1	V2	V1	V2	V1	V2
5	1	V1	V2	V1	V2	V1	V2
	2	V2	V1	V2	V1	V2	V1
6	1	V2	V1	V2	V1	V2	V1
	2	V1	V2	V1	V2	V1	V2
7	1	V1	V2	V1	V2	V1	V2
	2	V2	V1	V2	V1	V2	V1
8	1	V2	V1	V2	V1	V2	V1
	2	V1	V2	V1	V2	V1	V2
9	1	V1	V2	V1	V2	V1	V2
	2	V2	V1	V2	V1	V2	V1
10	1	V2	V1	V2	V1	V2	V1
	2	V1	V2	V1	V2	V1	V2

TABLE 3. Error Rates of Word Groups

Overall Error Rate	= 8.75%
Control Error Rate	= 8.72%
Original Error Rate	= 8.67%
Replacement Error Rate	= 8.90%
Misrecognitions	= 2.00%
Control	= 1.74%
Original	= 1.93%
Replacement	= 2.63%
Nonrecognitions	= 6.75%
Control	= 6.98%
Original	= 6.74%
Replacement	= 6.27%

TABLE 4. Word Pairs With Significantly Different Error Rates
(Total Number of Trials Per Word Group = 180)
($p(\chi^2)$, $p < .05$)

OVERALL ERRORS

ORIGINAL	# ERRORS	REPLACEMENT	# ERRORS
BURF	53	BRF	10
DELTA	34	DAIRYLAND	16
SOUTHERN	21	MOUNTAIN	54
DOWN ARROW	11	SCROLL DOWN	24
E4	13	EVENT 4	27
E9	6	EVENT 9	15
E10	5	EVENT 10	16
DELETE	44	REMOVE	17
FAN	33	RANGE	7

MISRECOGNITION ERRORS

ORIGINAL	# ERRORS	REPLACEMENT	# ERRORS
KNOWN	5	IDENTIFIED	17
UNKNOWN	1	UNIDENTIFIED	14
FINISH	9	OVER	0
SEND	7	SUBMIT	0
LEFT ARROW	7	SCROLL LEFT	1
E4	0	EVENT 4	11
E5	3	EVENT 5	10
E8	13	EVENT 8	29
E9	2	EVENT 9	10
E10	1	EVENT 10	8
LOCATE	0	COORDINATES	4
SUSPECT SITES	5	POSSIBLE SITES	0
DELETE	6	REMOVE	0
FAN	6	RANGE	0

NONRECOGNITION ERRORS

ORIGINAL	# ERRORS	REPLACEMENT	# ERRORS
KNOWN	32	IDENTIFIED	13
UNKNOWN	24	UNIDENTIFIED	6
FINISH	9	OVER	25
BURF	53	BRF	8
DELTA	33	DAIRYLAND	15
SOUTHERN	11	MOUNTAIN	45
DOWN ARROW	9	SCROLL DOWN	24
DELETE	38	REMOVE	17
FAN	19	RANGE	7

(best word underlined)

TABLE 5. Words With Overall Error Rates Greather Than 10%

Overall Error Rate	Words
11%	INDISTINCT, SOUTHRICH, SHOW, INTELLIGENCE REPORTS, ADS GSF, UNIDENTIFIED
12%	SOUTHERN, LEFT ARROW, SUPPRESS
13%	EDIT, SCROLL DOWN
14%	NO, CENTRAL, WEST, UNKNOWN, TWO, NINE, E8, OVER, SCROLL UP, SCROLL LEFT
15%	OUTPUT FORMAT, EVENT 4
17%	SEVEN, IDENTIFIED
18%	FAN
19%	YES, DELTA
20%	FOUR, EVENT 8
21%	NORTH, SOUTH, OUTT, KNOWN, EIGHT
22%	INN
24%	DELETE
29%	BURF
30%	MOUNTAIN

TOTAL = 39 Words

TABLE 6. Phonetic Groups in Vocabulary

MONOSYLLABLES	FINAL OR VESTIGAL T	BOTH
*YES	ENTER	*OUTT
*NO	*INDISTINCT	*WEST
*NORTH	RIVERTON	*EIGHT
*SOUTH	*DELTA	TEST
*INN	RIGHT ARROW	
*KNOWN	AUTO	TOTAL = 4
ONE	*EDIT	
*TWO	*E8	
THREE	LOCATE	
*FOUR	*OUTPUT FORMAT	
FIVE	SUSPECT SITES	
SIX	*DELETE	
*NINE	*IDENTIFIED	
*BURF	*UNIDENTIFIED	
CLEAR	*SCROLL LEFT	
SEND	SCROLL RIGHT	
*SHOW	EVENT 1	
LOG	*EVENT 8	
ALL	*MOUNTAIN	
*FAN	SUBMIT	
RANGE		
	TOTAL = 20	
TOTAL = 21		

*over 10% error rate

TABLE 7. Comparison of Error Rates of Phonetic Groups

Word Group	Total #	# With < 10% Errors		# With > 10% Errors		χ^2	Sig.
		Act.	Exp.	Act.	Exp.		
Monosyllables	21	9	14.4	12	6.55	6.44	<.05
Final or Vestigial T	20	9	13.8	11	6.44	5.28	<.05
Both	4	1	2.75	3	1.25	3.56	-
Neither	80	67	55.0	13	25.0	8.38	<.005
TOTAL	125	86		39		-	-

Other effects of phonetic structure on word recognition probably exist which were not detectable in this vocabulary, lacking as it is in size and diversity. Further research into the phonetic factors that affect machine intelligibility is needed.

REFERENCES

1. Crane, Peter M., Human Factors Comparison of Touch Screen and Voice Command Data Entry on a C3 System, SID, Digest of Technical Papers, 1984.

2. Yellen, Howard W., A Preliminary Analysis of Human Factors Affecting the Recognition Accuracy of a Discrete Word Recognizer for C3 Systems, Masters Thesis, Naval Postgraduate School, Monterey, California, 1983, 30-31.

3. Duddington, George R., and Thomas B. Schalk, "Speech Recognition: Turning Theory to Practice," IEEE Spectrum, September 1981, pp. 26-32.

BIBLIOGRAPHY

Levinson, Stephen E., and Mark Y. Liberman, "Speech Recognition by Computer," Scientific American, April 1981, pp. 64-73.

Naval Postgraduate School Report NPS55-81-013, A Longitudinal Study of Computer Voice Recognition Performance and Vocabulary Size, by Gary K. Poock, June 1981, 17.

Naval Postgraduate School Report NPS55-81-016, Effect of Operator Mental Loading on Voice Recognition System Performance, by J. W. Armstrong and Gary K. Poock, August 1981, 33.

SAS Institute, Inc., SAS User's Guide, Statistics, 1982 Edition, Cary: SAS Institute, 1982.

U S GOVERNMENT PRINTING OFFICE 1988 - 559-006/10235

END

FILMED

5-85

DTIC